AD-A049 622 STANFORD UNIV CALIF DEPT OF STATISTICS F/G 12/1
PRINCIPAL COMPONENTS IN THE NONNORMAL CASE: THE TEST FOR SPHERI--ETC(U)
OCT 77 C M WATERNAUX N00014-75-C-0442
UNCLASSIFIED TR-31 NL

1 OF 1
ADA049622

END
DATE FILMED
3-78
DDC

AD A049622

AD No.
DDC FILE COPY

12

6

PRINCIPAL COMPONENTS IN THE NONNORMAL CASE: THE TEST FOR SPHERICITY.

BY

10

CHRISTINE M. WATERNAUX

9

TECHNICAL REPORT, NO. 31

14

TR-31,
TR-122

11

OCTOBER 1977

12 28p.

15

PREPARED UNDER CONTRACT N00014-75-C-0442, NSF-MPS75-09450
(NR-042-034)
OFFICE OF NAVAL RESEARCH

THEODORE W. ANDERSON, PROJECT DIRECTOR

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

DDC
RECEIVED
FEB 8 1978
B

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

332 580

LB

PRINCIPAL COMPONENTS IN THE NONNORMAL CASE: THE TEST FOR SPHERICITY

BY

CHRISTINE M. WATERNAUX
Harvard University

TECHNICAL REPORT NO. 31
OCTOBER 1977

PREPARED UNDER CONTRACT N00014-75-C-0442
(NR-042-034)
OFFICE OF NAVAL RESEARCH

Theodore W. Anderson, Project Director

DDC
RECEIVED
FEB 8 1978
B

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Also issued as Technical Report No. 122 under National Science Foundation Grant MPS 75-09450 - Department of Statistics, Stanford University. Also, supported by National Science Foundation Grant MCS 76-09048.

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

PRINCIPAL COMPONENTS IN THE NONNORMAL CASE: THE TEST FOR SPHERICITY

Christine M. Waternaux
Harvard University

Summary.

The limiting distribution of the likelihood ratio statistic W_q for testing the hypothesis of equality of q characteristic roots of a covariance matrix for normal populations is studied for nonnormal populations. It is shown, both theoretically and empirically, that the limiting distribution of W_q is not robust to departures from normality characterized by nonzero fourth cumulants and that W_q cannot be used for these nonnormal populations. For the class of spherically symmetric populations, it is shown that the limiting distribution of W_q is proportional to a chi-square under the null hypothesis of equality of q population roots and to a noncentral chi-square under an appropriate sequence of alternative hypotheses. A corrected test statistic, W_q^*, whose limiting distribution is chi-square, is proposed to test the hypothesis of sphericity for the general class of spherically symmetric populations. Results of a Monte Carlo experiment conducted to compare the performances of W_q and W_q^* are presented for various contaminated normal models at various sample sizes. It is demonstrated that there is little difference between W_q and W_q^* for normal populations, but that dramatic improvements are gained by using W_q^* for the nonnormal populations.

KEY WORDS: Principal Components. Robustness. Test for Sphericity. Test for Equality of Variances.

for
White Section
Buff Section
JUSTIFICATION
BY
DISTRIBUTION/AVAILABILITY CODES
Dist. SPECIAL
A

1. Introduction. Principal Components Analysis is a multivariate data analysis technique which aims at reducing the dimensionality of a p-variate data set without losing too much information. The test for sphericity is the critical test which is performed in order to separate (p-q) principal components with large variances which significantly contribute to the total variation in the sample, from q components with much smaller and equal variances which explain a negligible fraction of the total variation.

Let $\underset{\sim}{x}$ be a $p \times 1$ random vector. Let $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p$ be the characteristic roots of the population matrix Σ, and let $\ell_1 < \ell_2 < \ldots < \ell_p$ denote the roots of the sample covariance matrix $S = (s_{ij})$ for a sample of size N drawn from the distribution of $\underset{\sim}{x}$. The vector $\underset{\sim}{y}$ of principal components is obtained by rotation of the original vector $\underset{\sim}{x}$: $\underset{\sim}{y} = \Gamma'\underset{\sim}{x}$, where Γ denotes the orthogonal matrix of (ordered) characteristic vectors of Σ. Then

$$\text{var } y_i = \lambda_i \qquad (i = 1, \ldots, p) ,$$

and the hypothesis of equality of q roots is:

$$H_0: \lambda_1 = \lambda_2 = \ldots = \lambda_q = \lambda < \lambda_{q+1} < \ldots < \lambda_p .$$

If $\underset{\sim}{x}$ has a multivariate normal distribution, the likelihood ratio criterion for testing H_0 is

$$\Lambda_q = \left\{ \prod_{i=1}^{q} \ell_i \Big/ \left(\frac{1}{q} \sum_{i=1}^{q} \ell_i \right)^q \right\}^{N/2}$$

and the limiting distribution for large N of $W_q = -2 \log \Lambda_q$ is chi-square on $(q+2)(q-1)/2$ degrees of freedom (Anderson, 1963).

The purpose of this paper is to show that the limiting distribution of W_q is nonrobust to departures from normality in the population, both under the null hypothesis and under the alternative, and to propose a modified test statistic W_q^* that can be used to test H_0 for the more general class of affine transformations of spherically symmetric populations (which include the normal populations).

The nonrobustness of the tests for the hypothesis of homogeneity of variances for p independent samples has been pointed out by Box (1953) and alternative methods have been proposed by several authors: Levene (1960), Miller (1968), Layard (1973) and Brown and Forsythe (1974). The testing problem considered here for the purpose of Principal Components Analysis is a generalization in the following sense: the p variables $x_i (i = 1, \ldots, p)$ are not necessarily independent and one tests the homogeneity of variances of the uncorrelated variables $y_i (i = 1, \ldots, q; q \leq p)$ obtained by an appropriate rotation of the x_i's. The rotation-invariant test statistic W_q is a function only of the sample roots $\ell_i (i = 1, \ldots, q)$. The limiting distribution of W_q as $N \to +\infty$ depends on the kurtosis of the marginal distribution of y_i and also on the bivariate fourth order cumulants, $\mathrm{cor}(y_i^2, y_j^2)$, as does the limiting distribution of the sample roots (Waternaux, 1976). In the general case, adjusting for nonzero fourth cumulants by estimating them is difficult and the limiting distributions of the corrected test statistics are not simple. However, for the class of spherically symmetric distributions - the class for which Principal Components Analysis is most appropriate and meaningful - a robust statistic W_q^*, whose limiting distribution as $N \to +\infty$ is also chi-square, is proposed to test for

sphericity. For this class of distributions, the significance level and power of the tests are (asymptotically) valid and the standard results of Principal Components Analysis can be used (after adjustment for kurtosis).

2. Asymptotic Theory.

Assumptions and Notation: Let $\underset{\sim}{x}$ be a $(p \times 1)$ vector from a multivariate distribution with $E(\underset{\sim}{x}) = 0$ and for which all fourth order moments and crossmoments exist and denote the standardized cumulants by $\kappa_{1111}^{ij\ell m}$ $(i, j, \ell, m = 1, \ldots, p)$; for example

$$\kappa_4^i = \frac{E[x_i^4]}{(\text{var } x_i)^2} - 3 \qquad (i = 1, \ldots, p) .$$

$$\kappa_{22}^{ij} = \frac{E[x_i^2, x_j^2]}{\text{var } x_i \text{ var } x_j} - 1 \qquad (i \neq j) .$$

For the asymptotic theory assume (without loss of generality) that

$$\Sigma = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p) ,$$

and that the $(p-q)$ largest roots are distinct

$$\lambda_{q+1} < \lambda_{q+2} < \ldots < \lambda_p . \tag{1}$$

The following standardized variables are constructed from the elements of the sample covariance matrix S

$$z_{ij} = \sqrt{N/\lambda_i\lambda_j}\,(s_{ij} - \delta_{ij}\lambda_i) \qquad (i,j = 1, \ldots, p) , \tag{2}$$

where δ_{ij} is the Kronecker delta; let $Z = (z_{ij})$.

The probability law of a random variable X will be indicated by $\mathcal{L}(X)$. The Mann-Wald symbols O_p and o_p are interpreted in the usual sense. Two lemmas will be useful.

Lemma 1 (Slutsky Theorem). Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables. If $X_n = Y_n + o_p(1)$ and $\mathcal{L}(Y_n) \to \mathcal{L}(Y)$ as $n \to +\infty$, then $\mathcal{L}(X_n) \to \mathcal{L}(Y)$ as $n \to +\infty$.

As a consequence, if $Y_n \xrightarrow{p} a$, then $\mathcal{L}\left(\frac{X_n}{Y_n}\right) \to \mathcal{L}\left(\frac{X}{a}\right)$ as $n \to +\infty$ since $\mathcal{L}\left(\frac{X_n}{Y_n}\right) = \mathcal{L}\left(\frac{X_n}{a}\,[1 + o_p(1)]\right)$.

Lemma 2 (see, for instance, Rao [1965]). Let $\underset{\sim}{x}$ be multivariate normal with mean $\underset{\sim}{\mu}$ and nonsingular covariance matrix V. A necessary and sufficient condition for $\underset{\sim}{x}'A\underset{\sim}{x}$ to be noncentral chi-square with r degrees of freedom and noncentrality parameter $\frac{1}{2}\,\underset{\sim}{\mu}'A\underset{\sim}{\mu}$ is that AV be idempotent and $r = \operatorname{rank}(AV)$.

The asymptotic distributions of W_q under the null and alternative hypotheses are now found by first deriving an asymptotic expansion in terms of the standardized variables z_{ij}.

2.1. Asymptotic Expansion. The test criterion

$$W_q = -N\left\{\log \prod_{i=1}^{q} \ell_i - q \log\left(\frac{1}{q}\sum_{i=1}^{q} \ell_i\right)\right\}$$

is rewritten as

$$N\left\{q \log\left(\operatorname{tr}S - \sum_{i=q+1}^{p} \ell_i\right) - \log|S| + \sum_{i=q+1}^{p} \log \ell_i\right\}. \tag{3}$$

The following expansions hold for the determinant and trace of S:

$$\log|S| = \log|\Lambda| + \frac{\operatorname{tr}Z}{\sqrt{N}} - \frac{\operatorname{tr}Z^2}{2N} + O_p(N^{-3/2}),$$

$$\operatorname{tr}S = \operatorname{tr}\Lambda + \frac{\operatorname{tr}\Lambda Z}{\sqrt{N}}.$$

It is known that when the population root λ_r is simple the rth sample root ℓ_r can be expanded as:

$$(4) \qquad \ell_r = \lambda_r \left\{ 1 + \frac{z_{rr}}{\sqrt{N}} + \frac{1}{N} \sum_{i \neq r} \frac{\lambda_i z_{ri}^2}{\lambda_r - \lambda_i} + O_p(N^{-3/2}) \right\} ,$$

(see Lawley, 1956); this holds for $r = q+1, \ldots, p$ under the assumption (1).

Substituting in (3) the expansion (4) for ℓ_r and the expansions for $\log|S|$ and trS, some straightforward algebra yields the asymptotic expansion for W_q

$$(5) \qquad W_q = -N \sum_{k=1}^{q} \log \rho_k + \sqrt{N} \sum_{k=1}^{q} (\rho_k - 1) z_{kk} + \sum_{\substack{j,k=1 \\ j<k}}^{q} z_{jk}^2$$

$$+ \sum_{k=1}^{q} \sum_{i=q+1}^{p} \frac{\lambda_k(\rho_k - 1)}{\lambda_i - \lambda_k} z_{ik}^2 + \frac{1}{2} \sum_{k=1}^{q} z_{kk}^2 - \frac{1}{2q} \left(\sum_{k=1}^{q} \rho_k z_{kk} \right)^2$$

$$+ O_p(N^{-1/2}) ,$$

where $\rho_k = \lambda_k / \bar{\lambda}$, and $\bar{\lambda} = \frac{1}{q} \sum_{j=1}^{q} \lambda_j$.

The expansion (5) is valid whether or not the null hypothesis H_0 is satisfied provided that $\lambda_{q+1} < \lambda_{q+2} < \ldots < \lambda_p$. This expansion does not seem to have been published in the literature. Under the null hypothesis, $\lambda_1 = \ldots = \lambda_q = \bar{\lambda} = \lambda$, the expansion of W_q reduces to

$$(6) \qquad W_q = \sum_{\substack{i,j=1 \\ i<j}}^{q} z_{ij}^2 + \frac{1}{2} \sum_{i=1}^{q} (z_{ii} - \bar{z})^2 + O_p(N^{-1/2}) ,$$

where

$$\bar{z} = \frac{1}{q} \sum_{i=1}^{q} z_{ii} .$$

The latter expansion has been given by Anderson (1963), using another method.

Note that the departures from sphericity are decomposed into two parts in the expansion of W_q: $q(q-1)/2$ off diagonal terms indicating correlation and $(q-1)$ diagonal terms indicating heterogeneity of variance.

It should be noted that under the null hypothesis $\lambda_1 = \ldots = \lambda_q = \lambda$ the asymptotic expansion of W_q, up to terms of order $N^{-1/2}$, only involves z_{ij} where $i \leq j \leq q$ and therefore the limiting distribution of W_q only depends on the distribution of $y_i (i = 1, \ldots, q)$.

2.2. Limiting Distribution of W_q Under the Null Hypothesis.

Proposition 1. The limiting null distribution as $N \to +\infty$ of W_q is that of a linear combination of chi-squares on one degree of freedom whose coefficients depend on the (standardized) fourth cumulants of the principal components of the parent population

$$\kappa_4^i \ (i = 1, \ldots, q) \quad \text{and} \quad \kappa_{22}^{ij} \ (i,j = 1, \ldots, q; \ i \neq j) \ .$$

For normal populations (or for populations whose fourth cumulants are zero), the limiting null distribution of W_q is chi-square on $(q+2)(q-1)/2$ degrees of freedom.

Proof. By the multivariate central limit theorem the variables z_{ij} are asymptotically jointly multivariate normal with covariance matrix given by

$$(7) \qquad \operatorname{var} z_{ii} = \kappa_4^i + 2 + 0(N^{-1}) \qquad (i = 1, \ldots, p) \ ,$$

$$\operatorname{var} z_{ij} = \kappa_{22}^{ij} + 1 + 0(N^{-1}) \qquad (i,j = 1, \ldots, p; \ i \neq j) \ ,$$

$$\operatorname{cov}(z_{ij}, z_{\ell m}) = \kappa_{1111}^{ij\ell m} + 0(N^{-1}) \qquad (i,j) \neq (\ell,m) \ .$$

For normal populations the z_{ij}'s are asymptotically independent, and the well known result

$$\lim_{N\to+\infty} \mathcal{L}(W_q) = \chi^2[(q+2)(q-1)/2]$$

follows immediately from (6) and Slutsky's theorem. For nonnormal populations the z_{ij}'s are in general dependent and

$$\lim_{N\to+\infty} \mathcal{L}(z_{ii}^2) = (\kappa_4^i + 2)\ \chi^2(1)\ ,$$

$$\lim_{N\to+\infty} \mathcal{L}(z_{ij}^2) = (\kappa_{22}^{ij} + 1)\ \chi^2(1)\ ,$$

and, in general, the limiting distribution of W_q is complicated.

Finally, if the observations are a random sample of p independent identically distributed variables x_i with standardized kurtosis κ_4, the limiting distribution of W_q is that of a linear combination of two independent chi-squares

$$\lim_{N\to+\infty} \mathcal{L}(W_q) = \chi^2[q(q-1)/2] + (1 + \kappa_4/2)\ \chi^2(q-1)\ .$$

2.3. Limiting Distribution Under Alternative Hypotheses.

Proposition 2. Under the alternative hypothesis ($\rho_i = \lambda_i/\bar{\lambda} \neq 1$ for some $i = 1, \ldots, q$) the limiting distribution as $N \to +\infty$ of

$$(W_q + N \sum_{j=1}^{q} \log \rho_j)/\sqrt{N}$$

is normal with mean 0 and variance

$$\sum_{j=1}^{q} (\rho_j - 1)^2\ (\kappa_4^j + 2) + 2 \sum_{\substack{i,j=1 \\ i<j}}^{q} \rho_i \rho_j\ \kappa_{22}^{ij}\ .$$

If the standardized variables $\sqrt{\lambda_i}\, x_i$ $(i = 1, \ldots, p)$ are independent and identically distributed with kurtosis κ_4, the variance reduces to

$$(\kappa_4 + 2) \sum_{j=1}^{q} (\rho_j - 1)^2 .$$

Proof. The result follows from the asymptotic expansion (5) for W_q.

Proposition 3. (a) Under the sequence of alternatives

$$H_{aN}: \quad \lambda_i = \lambda\left(1 + \frac{a_i}{\sqrt{N}}\right) \qquad (\forall i=1, \ldots, q) ,$$

where $\underset{\sim}{a} \in R^q$ and $\lambda_i > 0$, the limiting distribution for large N of W_q is that of a linear combination of central and noncentral chi-squares on one degree of freedom. In the particular case of a random sample of $\underset{\sim}{x}$, where the variables $\sqrt{\lambda_i}\, x_i$ are independent and identically distributed with kurtosis κ_4, the limiting distribution of W_q is that of a linear combination of a central chi-square and an independent noncentral chi-square

$$\lim_{N \to +\infty} \mathcal{L}(W_q) = \chi^2[q(q-1)/2] + (1 + \kappa_4/2)\, \chi'^2_f(q-1) ,$$

where the noncentrality parameter is

$$f = \sum_{i=1}^{q} (a_i - \bar{a})^2/(4 + 2\kappa_4) .$$

(b) Under the sequence of alternatives

$$H'_{aN}: \quad \lambda_i = \lambda\left(1 + \frac{a_i}{N}\right) \qquad (\forall i = 1, \ldots, q) ,$$

the limiting distribution as $N \to +\infty$ of W_q is the same as in the null case.

Proof. The result follows from the fact that in the case (a), under H_{aN}

$$\rho_i = \frac{\lambda_i}{\bar{\lambda}} = 1 + \frac{a_i - \bar{a}}{\sqrt{N}} + O(N^{-1}) \quad (i = 1, \ldots, q) ,$$

and after simplifications, the asymptotic expansion (5) is reduced to

$$W_q = \frac{1}{2} \sum_{i=1}^{q} (z_{ii} - a_i - \bar{z} - \bar{a})^2 + \sum_{\substack{i,j=1 \\ i<j}}^{q} z_{ij}^2 + O_p(N^{-1/2}) .$$

Similarly for (b), under H'_{aN}

$$\rho_i = \frac{\lambda_i}{\bar{\lambda}} = 1 + \frac{a_i - \bar{a}}{N} + O(N^{-2}) \quad (i = 1, \ldots, q) ,$$

and the constant terms in the expansion (5) are of order N^{-1}. The asymptotic expansion for W_q is then given by (6), as in the null case, and the limiting distribution of W_q is the same as under the null hypothesis.

3. Distribution Theory for Spherically Symmetric Models. Consider the class of p-variate distributions which are spherically symmetric after an appropriate affine transformation. More precisely, define $\mathfrak{F}$ by:

(8) the distribution of $\underset{\sim}{x}$, $F(\underset{\sim}{x}|\Sigma)$, $\in \mathfrak{F}$ iff

(i) All moments and crossmoments of $\underset{\sim}{x}$ or order ≤ 4 exist and the covariance matrix Σ is nonsingular. Let $E(\underset{\sim}{x}) = 0$ without loss of generality and let $\text{var}(\underset{\sim}{x}) = \Sigma = \Gamma\Lambda\Gamma'$, as previously.

(ii) The distribution of $\underset{\sim}{w} = \Sigma^{-1/2} \underset{\sim}{x}$ is spherically symmetric, that is: for all orthogonal $(p \times p)$ matrix H, $\underset{\sim}{w}$ and $H\underset{\sim}{w}$ have the same distribution.

If the density exists, it can be expressed in the form

$$|\Sigma|^{-1/2} g(\underset{\sim}{x}' \Sigma^{-1} \underset{\sim}{x})$$

for some function g. (See for instance Lord, 1954 and Kelker, 1970.)

The class $\mathfrak{F}$ provides a simple yet quite large class of multivariate models as an alternative to multivariate normal models. Note that the standardized kurtosis κ_4^i is the same for all marginal distributions $(i = 1, \ldots, p)$. By varying g one can generate both long-tailed $(\kappa_4^i > 0)$ or short-tailed $(\kappa_4^i < 0)$ models; this is particularly appropriate for the present study since the limiting distributions depend only on the fourth cumulants of the parent population. The class $\mathfrak{F}$ includes in particular the p-variate normal model $\Phi(\underset{\sim}{x}|\Sigma)$ and the ε-contaminated normal models with distribution function given by

$$\Phi_\varepsilon(\underset{\sim}{x}|\Sigma,\alpha) = (1 - \varepsilon)\, \Phi(\underset{\sim}{x}|\Sigma) + \varepsilon H(\underset{\sim}{x}|\alpha\Sigma) ,$$

where $0 < \varepsilon < 1$, α is a positive scalar and $H(\underset{\sim}{x}|\Sigma) \in \mathfrak{F}$; it also includes the multivariate T distribution.

It will be shown that considerable simplifications appear in the limiting distributions of W_q for any population in $\mathfrak{F}$. These simplifications arise from the fact that the standardized cumulants of the principal components y_i $(i = 1, \ldots, p)$ satisfy then

$$(9) \qquad \kappa_4^i = 3\,\kappa_{22}^{ij} = 3\kappa, \quad \text{say}, \quad (\forall i,j = 1, \ldots, p;\ i \neq j)\ ,$$

$$\kappa_{31}^{ij} = \kappa_{1111}^{ij\ell m} = 0 \qquad (\forall i \neq j \ \text{ or } \ \ell \neq m)\ .$$

3.1. Asymptotic Distributions of W_q

Proposition 4. For any population $F(\underset{\sim}{x}|\Sigma) \in \mathfrak{F}$, under the null hypothesis

$$H_0: \ \lambda_1 = \lambda_2 = \ldots = \lambda_q = \lambda < \lambda_{q+1} < \ldots < \lambda_p$$

the limiting distribution of $W_q/(\kappa+1)$ as $N \to +\infty$ is chi-square on $(q+2)(q-1)/2$ degrees of freedom.

Proof. As for the general case, the limiting distribution under H_0 of W_q is that of

$$T = \sum_{\substack{i,j=1 \\ i<j}}^{q} z_{ij}^2 + \frac{1}{2}\sum_{i=1}^{q} (z_{ii} - \bar{z})^2\ .$$

Consider

$$\underset{\sim}{u} = \left(\frac{z_{11}}{\sqrt{2}}, \ \ldots, \ \frac{z_{qq}}{\sqrt{2}}, \ z_{21}, \ z_{31}, \ z_{32}, \ \ldots, \ z_{q,(q-1)}\right)/\sqrt{\kappa+1}$$

then

$$T = (\kappa+1)\ \underset{\sim}{u}'H\underset{\sim}{u}$$

where H is the block diagonal matrix

$$H = \begin{pmatrix} I_q - \dfrac{ee'}{q} & 0 \\ 0 & I_{q(q-1)/2} \end{pmatrix}$$

(I_q denotes the $q \times q$ identity matrix and $\underset{\sim}{e}' = (1, \ldots, 1) \in R^q$.) The limiting distribution of $\underset{\sim}{u}$ is normal with mean 0 and covariance matrix

$$V = \begin{pmatrix} I_q + \frac{\kappa}{2\kappa+2} \underset{\sim}{e}\underset{\sim}{e}' & 0 \\ 0 & I_{q(q-1)/2} \end{pmatrix}$$

by substituting $\kappa_4^i = 3\kappa_{22}^{ij} = \kappa$ in (7). Then, by Lemma 2,

$$\lim_{N \to +\infty} \mathcal{L}(\underset{\sim}{u}'H\underset{\sim}{u}) = \chi^2(r)$$

since V is nonsingular and VH is idempotent of rank $r = (q+2)(q-1)/2$, and Proposition 3 holds.

Proposition 5. (a) For any distribution $F \in \mathfrak{F}$, under the sequence of alternatives

$$H_{aN}: \quad \lambda_i = \lambda\left(1 + \frac{a_i}{\sqrt{N}}\right) \qquad (\forall i = 1, \ldots, q) ,$$

where $\underset{\sim}{a} \in R^q$ and $\lambda_i > 0$, the limiting distribution as $N \to +\infty$ of $W_q/(\kappa+1)$ is noncentral chi-square on $(q+2)(q-1)/2$ degrees of freedom and noncentrality parameter

$$f = \sum_{i=1}^{q} (a_i - \bar{a})^2/4(\kappa+1), \quad \text{where} \quad \bar{a} = \frac{1}{q} \Sigma\, a_i .$$

(b) For any distribution $F \in \mathfrak{F}$, under the sequence of alternatives

$$H'_{aN}: \quad \lambda_i = \lambda\left(1 + \frac{a_i}{N}\right) \qquad (\forall i = 1, \ldots, q)$$

the limiting distribution as $N \to +\infty$ of $W_q/(\kappa+1)$ is chi-square on $(q+2)(q-1)/2$ degrees of freedom.

Proof. As in the proof of Proposition 4, note that in the case (a) the limiting distribution of W_q is that of

$$\frac{1}{2}\sum_{i=1}^{q}(z_{ii} - a_i - \bar{z} - \bar{a})^2 + \sum_{\substack{i,j=1 \\ i<j}}^{q} z_{ij}^2 ,$$

and in the case (b) the same as under the null hypothesis. Then for (a) consider

$$v = \left(\frac{z_{11} - a_1}{\sqrt{2}}, \ldots, \frac{z_{qq} - a_q}{\sqrt{2}}, z_{21}, z_{31}, z_{32}, \ldots, z_{q,(q-1)}\right)/\sqrt{\kappa+1}$$

the limiting distribution of $\underset{\sim}{v}$ is normal with mean

$$\underset{\sim}{\delta}' = \left(-\frac{a_1}{\sqrt{2}}, \ldots, -\frac{a_q}{\sqrt{2}}, 0, 0, \ldots, 0\right)/\sqrt{\kappa+1}$$

and covariance matrix V. Then, by Lemmas 1 and 2,

$$\lim_{N\to+\infty} \mathcal{L}\left(\frac{W_q}{\kappa+1}\right) = \lim_{N\to+\infty} \mathcal{L}(\underset{\sim}{v}'H\underset{\sim}{v}) = \chi_f'^2(r) .$$

Proposition 5 generalizes two results given by Nagao (1970) for the limiting distribution of W_q when $q = p$ and the population is multivariate normal. By expanding the characteristic function of $-2 \log \Lambda_p$, Nagao showed that under the sequence of alternatives

$$K_{N,\delta}: \quad \Sigma = I + \frac{\Omega}{N^{\delta}}$$

(where Ω is a symmetric positive matrix), the limiting distribution of W_p is noncentral chi-square $\chi_f'^2(r)$, with noncentrality parameter

$$f = \frac{1}{4}[\mathrm{tr}\Omega^2 - \frac{1}{p}(\mathrm{tr}\Omega)^2]$$

when $\delta = \frac{1}{2}$, and is central chi-square on r degrees of freedom when $\delta = 1$. For spherically symmetric models under the sequence of alternatives

$$\Sigma = \lambda\left(I + \frac{\Omega}{N^{\delta}}\right)$$

the limiting distribution of $W_p/(\kappa+1)$ is noncentral chi-square $\chi_f'^2(r)$, where

$$f = \frac{1}{4(\kappa+1)}\left[\text{tr } \Omega^2 - \frac{1}{p}(\text{tr } \Omega)^2\right]$$

when $\delta = \frac{1}{2}$, and central chi-square on r degrees of freedom when $\delta = 1$.

3.2. Robust Test for Sphericity. The previous asymptotic distributions suggest a reasonable method for testing for sphericity in the class $\mathfrak{F}$ of spherically symmetric models.

Proposition 6. For any population in the class $\mathfrak{F}$ defined by (8), the test statistic

$$W_q^* = \frac{W_q}{\hat{\kappa}+1},$$

where $\hat{\kappa}$ is a consistent estimator of κ, is asymptotically chi-square on $r = (q+2)(q-1)/2$ degrees of freedom under the null hypothesis H_0 and the sequence of alternatives $\{H'_{aN}\}$ and noncentral chi-square under the sequence of alternatives $\{H_{aN}\}$.

Proof. The results follow from Lemma 1 and Propositions 3 and 4.

The asymptotic null distribution of W_q^* is the same for all populations in $\mathfrak{F}$ and therefore W_q^* can be used to test the hypothesis of sphericity.

Estimation of κ. Under the null hypothesis $\lambda_i = \lambda$ $(i = 1, \ldots, q)$, one notes that

$$E(R^2) = E\left(\sum_{i=1}^{q} x_i^2\right) = q\lambda$$

$$E(R^4) = E\left(\sum_{i=1}^{q} x_i^2\right)^2 = q(q+2)(\kappa+1)\ \lambda^2 ,$$

where it is recalled that the x_i's are assumed to have no correlation. Thus the nuisance parameter κ satisfies

$$(\kappa+1) = \frac{q}{q+2} \frac{E(R^4)}{(E(R^2))^2}$$

and a consistent estimator of κ is

$$\hat{\kappa} = \frac{q}{q+2} \frac{m_4}{(m_2)^2} - 1$$

where

$$m_2 = \frac{1}{N} \sum_{\alpha=1}^{N} \left(\sum_{i=1}^{q} y_{i\alpha}^2 \right) ,$$

$$m_4 = \frac{1}{N} \sum_{\alpha=1}^{N} \left(\sum_{i=1}^{q} y_{i\alpha}^2 \right)^2 ,$$

and $y_{i\alpha}$ is the ith component of the αth observation. The sample components are used because, in practice, the population covariance matrix is not necessarily diagonal. Note that if $p = q$

$$\sum_{i=1}^{q} y_{i\alpha}^2 = \sum_{i=1}^{q} x_{i\alpha}^2 ,$$

and it is not necessary to use the $y_{i\alpha}$'s to estimate κ.

The following points are worth raising about the estimation of the kurtosis of a distribution in general, and κ in particular, for long-tailed populations.

(i) The proposed estimate $\hat{\kappa}$ is not "robust" in the sense that its influence curve (see Hampel, 1973) is not bounded, and therefore one can expect sensitivity to large outliers. However κ is actually a standardized kurtosis and a large observation will inflate both numerator (m_4) and denominator (m_2^2). Also $\hat{\kappa}$ is, in this case, computed by summing over q dimensions, and turned out to be considerably more stable than the estimate of the kurtosis of a single variable.

(ii) Numerous attempts have been made by the author to produce a robust estimate of kurtosis such as winsorizing or weighting down large observations, and several estimates were empirically studied using Monte Carlo methods. The robustified estimates exhibited a considerable bias towards zero, making the correction term $(1+\hat{\kappa})$ ineffectual.

(iii) Finally, further effort in the estimation of κ was abandoned when the empirical results showed that the performances of $W_q/(\kappa+1)$, where the true value for κ was substituted for each population (corresponding to an hypothetical perfect estimate), were not superior to the performances of W_q^*. It seems that the presence of a large outlier increases W_q and $\hat{\kappa}$ simultaneously, but does not affect W_q^* as severely.

3.3. Generalization to a Wider Class of Distributions. It should be emphasized that the previous results hold for an even wider class of populations than the class $\mathfrak{F}$. It was remarked in Section 2.1 that under the null hypothesis (or under a sequence of alternatives close enough to the null hypothesis) the limiting distribution of W_q is determined by the limiting distribution of z_{ij}, where $i,j = 1, \ldots, q$

only. More precisely, the proofs for Propositions 3 and 4 depend only on the assumption that the asymptotic covariance matrix of z_{ii} $(i = 1, \ldots, q)$ has a structure of an intraclass correlation matrix, that is, of the form

$$(aI_q + b\underset{\sim}{e}\underset{\sim}{e}')$$

and that the asymptotic covariance of z_{ij}

$$\text{var } z_{ij} = \frac{1}{2}(a-b) \qquad (i,j = 1, \ldots, q;\ i \neq j)$$

or, equivalently, on the assumption that the standardized cumulants of the population principal components satisfy (9). In summary:

Proposition 7. Let $\underset{\sim}{x}$ be a $(p \times 1)$ vector for which all moments of order 4 exist. Let $\Sigma = \Gamma\Lambda\Gamma'$ denote the covariance matrix of $\underset{\sim}{x}$. A sufficient condition for the limiting null distribution of $W_q/(\kappa+1)$ to be chi-square on r degrees of freedom is that

$$\text{var } y_i = \lambda \qquad (i = 1, \ldots, q)\ ,$$

where $\underset{\sim}{y} = \Gamma'\underset{\sim}{x}$ is the $(p \times 1)$ vector of the (ordered) principal components, and that the standardized cumulants of $\underset{\sim}{y}$ satisfy

$$\kappa_4^i = 3\,\kappa_{22}^{ij} \qquad (i,j=1, \ldots, q;\ i \neq j)\ ,$$

$$\kappa_{1111}^{ijkm} = 0 \qquad (i,j,k,m \leq q;\ i \neq j \text{ or } k \neq m)\ .$$

In particular, it is sufficient that the joint distribution of $(y_1, \ldots, y_q)$ is spherically symmetric.

4. Monte Carlo Experiment. A Monte Carlo experiment has been conducted to study the sampling distributions of the test statistics W_q and

W_q^* for variouspopulations and sample sizes $N = 50, 100, 200$ under the null and alternative hypotheses. The different populations considered were

(i) a multivariate normal population,

(ii) a short-tailed population which is a mixture of two normal populations with different means,

and several contaminated normal populations with distribution function

$$\phi_\varepsilon(\underset{\sim}{x}|\Sigma,\alpha) = (1-\varepsilon)\ \Phi(\underset{\sim}{x}|\Sigma) + \varepsilon\Phi(\underset{\sim}{x}|\alpha\Sigma)$$

(where $\Phi(\underset{\sim}{x}|\Sigma)$ denotes the distribution function of a multivariate normal vector $\underset{\sim}{x}$ with covariance matrix Σ), with parameters:

(iii) $\varepsilon = .3 \quad \sqrt{\alpha} = 2$,

(iv) $\varepsilon = .1 \quad \sqrt{\alpha} = 3$.

The dimensions and population roots considered were

(i) $p = q = 2$ and $\lambda_1 = \lambda_2 = 1$,

(ii) $p = q = 6$ and $\lambda_1 = \lambda_2 = \ldots = \lambda_6 = 1$,

(iii) $p = 6$, $q = 4$ and $\lambda_1 = \ldots = \lambda_4 = 1$, $\lambda_5 = 3$, $\lambda_6 = 5$.

For each sample size 200 samples were drawn for each population and the corresponding values for both test statistics W_q and W_q^* were computed. The sampling distributions of W_q and W_q^* have then been compared to the limiting distributions given by the asymptotic theory: the central chi-square on r degrees of freedom under the null hypothesis, the normal approximation as well as the central and noncentral chi-square under the alternative.

The simulations showed that the asymptotic chi-square approximation to the distributions of $W_q/\kappa+1$ and W_q^* is good for $N = 100, 200$, but somewhat less accurate for $N = 50$; for normal populations, however,

the asymptotic approximation of the distribution of W_q by a chi-square is accurate for sample sizes as small as $N = 50$. Tables 1 and 2 present some numerical results of the simulations.

It is clearly demonstrated in Tables 1 and 2 that the Lawley-Bartlett criterion W_q cannot be used for testing the hypothesis of sphericity when the parent population is nonnormal, even for moderate departures from normality as the contaminated normal $\epsilon = .3$, $\sigma = 2$ (the standardized kurtosis is then $3\kappa = 1.6$). As an example, for $N = 200$, $p = q = 6$, the observed significance level of the test based on W_q is 50%, for a nominal level of 5%. On the other hand, the test based on the proposed statistic W_q^* has an observed significance level of 7.5%, not significantly different from the nominal level of 5%. For a normal population, the test based on W_q is the likelihood ratio test but test criteria W_q and W_q^* are then almost identical, and therefore W_q^* can be used without loss of efficiency at the normal model.

For short-tailed models, as expected, the results of the simulations were not as spectacular. Since the kurtosis of any population should satisfy $\kappa_4 \geq -2$, no short-tailed model can produce the same extreme behavior for W_q as occurs for lont-tailed models. For moderately short-tailed models, the sampling distribution of W_q is well approximated by a chi-square.

TABLE 1. Number of times (out of 200) W_q and W_q^* exceed $\chi^2_\alpha(20)$
$p = q = 6$, $N = 200$

Model	α-level test	$\alpha = .25$	$\alpha = .05$	$\alpha = .01$
Normal	W	53	10	4
	W*	52	10	5
Cont. Normal $\varepsilon = .3$, $\sigma = 2$ ($3\kappa = 1.6$)	W	146	99	37
	W*	52	15	7
Cont. Normal $\varepsilon = .1$, $\sigma = 3$ ($3\kappa = 5.3$)	W	198	188	160
	W*	53	13	6
	exp. number†	50 ± 12	10 ± 6	2 ± 3

TABLE 2. Number of times (out of 200) W_q and W_q^* exceed $\chi^2_\alpha(9)$
$p = 6$, $q = 4$, $N = 200$

Model	α-level test	$\alpha = .25$	$\alpha = .05$	$\alpha = .01$
Normal	W	55	11	1
	W*	49	9	1
Cont. Normal $\varepsilon = .3$, $\sigma = 2$ ($3\kappa = 1.6$)	W	116	51	17
	W*	51	12	2
Cont. Normal $\varepsilon = .1$, $\sigma = 3$ ($3\kappa = 5.3$)	W	183	151	117
	W*	55	10	3
	exp. number†	50 ± 12	10 ± 6	2 ± 3

†expected number = $200\alpha \pm 2\sqrt{200\alpha(1-\alpha)}$

Bibliography

Anderson, T. W. (1963). Asymptotic theory for principal components. Ann. Math. Statist. 34, 122-148.

Box, G. E. P. (1953). Non-normality and tests on variances. Biometrika 40, 318-335.

Brown, M. B. and Forsythe, A. B. (1974). Robust tests for inequality of variances. J. Amer. Statist. Assoc. 69, 364-367.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. J. Amer. Statist. Assoc. 69, 383-393.

Kelker, D. (1970). Distribution theory of spherical distributions and a location scale parameter generalization. Sankhyā A 32, 419-430.

Lawley, D. N. (1956). Tests of significance for the latent roots of covariance and correlation matrix. Biometrika 43, 128-136.

Layard, M. W. J. (1973). Robust large-sample for homogeneity of variances. J. Amer. Statist. Assoc. 68, 195-198.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, ed., Contributions to Probability and Statistics: Stanford Univ. Press.

Lord, R. D. (1954). The use of Hankel transform in statistics. I. General theory and examples. Biometrika 41, 44-55.

Miller, R. G. (1968). Jackknifing variances. Ann. Math. Statist. 33, 567-582.

Nagao, H. (1970). Asymptotic expansions of some test criteria for homogeneity of variances and covariances matrices from normal populations. J. Sci. Hiroshima Univ. Ser. A I, 34, 153-247.

Waternaux, C. M. (1976). Asymptotic distribution of the sample roots for a nonnormal population. Biometrika 63, 639-645.

TECHNICAL REPORTS

OFFICE OF NAVAL RESEARCH CONTRACT N00014-67-A-0112-0030 (NR-042-034)

1. "Confidence Limits for the Expected Value of an Arbitrary Bounded Random Variable with a Continuous Distribution Function," T. W. Anderson, October 1, 1969.

2. "Efficient Estimation of Regression Coefficients in Time Series," T. W. Anderson, October 1, 1970.

3. "Determining the Appropriate Sample Size for Confidence Limits for a Proportion," T. W. Anderson and H. Burstein, October 15, 1970.

4. "Some General Results on Time-Ordered Classification," D. V. Hinkley, July 30, 1971.

5. "Tests for Randomness of Directions against Equatorial and Bimodal Alternatives," T. W. Anderson and M. A. Stephens, August 30, 1971.

6. "Estimation of Covariance Matrices with Linear Structure and Moving Average Processes of Finite Order," T. W. Anderson, October 29, 1971.

7. "The Stationarity of an Estimated Autoregressive Process," T. W. Anderson, November 15, 1971.

8. "On the Inverse of Some Covariance Matrices of Toeplitz Type," Raul Pedro Mentz, July 12, 1972.

9. "An Asymptotic Expansion of the Distribution of "Studentized" Classification Statistics," T. W. Anderson, September 10, 1972.

10. "Asymptotic Evaluation of the Probabilities of Misclassification by Linear Discriminant Functions," T. W. Anderson, September 28, 1972.

11. "Population Mixing Models and Clustering Algorithms," Stanley L. Sclove, February 1, 1973.

12. "Asymptotic Properties and Computation of Maximum Likelihood Estimates in the Mixed Model of the Analysis of Variance," John James Miller, November 21, 1973.

13. "Maximum Likelihood Estimation in the Birth-and-Death Process," Niels Keiding, November 28, 1973.

14. "Random Orthogonal Set Functions and Stochastic Models for the Gravity Potential of the Earth," Steffen L. Lauritzen, December 27, 1973.

15. "Maximum Likelihood Estimation of Parameters of an Autoregressive Process with Moving Average Residuals and Other Covariance Matrices with Linear Structure," T. W. Anderson, December, 1973.

16. "Note on a Case-Study in Box-Jenkins Seasonal Forecasting of Time series," Steffen L. Lauritzen, April, 1974.

TECHNICAL REPORTS (continued)

17. "General Exponential Models for Discrete Observations," Steffen L. Lauritzen, May, 1974.

18. "On the Interrelationships among Sufficiency, Total Sufficiency and Some Related Concepts," Steffen L. Lauritzen, June, 1974.

19. "Statistical Inference for Multiply Truncated Power Series Distributions," T. Cacoullos, September 30, 1974.

Office of Naval Research Contract N00014-75-C-0442 (NR-042-034)

20. "Estimation by Maximum Likelihood in Autoregressive Moving Average Models in the Time and Frequency Domains," T. W. Anderson, June 1975.

21. "Asymptotic Properties of Some Estimators in Moving Average Models," Raul Pedro Mentz, September 8, 1975.

22. "On a Spectral Estimate Obtained by an Autoregressive Model Fitting," Mituaki Huzii, February 1976.

23. "Estimating Means when Some Observations are Classified by Linear Discriminant Function," Chien-Pai Han, April 1976.

24. "Panels and Time Series Analysis: Markov Chains and Autoregressive Processes," T. W. Anderson, July 1976.

25. "Repeated Measurements on Autoregressive Processes," T. W. Anderson, September 1976.

26. "The Recurrence Classification of Risk and Storage Processes," J. Michael Harrison and Sidney I. Resnick, September 1976.

27. "The Generalized Variance of a Stationary Autoregressive Process," T. W. Anderson and Raul P.Mentz, October 1976.

28. "Estimation of the Parameters of Finite Location and Scale Mixtures," Javad Behboodian, October 1976.

29. "Identification of Parameters by the Distribution of a Maximum Random Variable," T. W. Anderson and S.G. Ghurye, November 1976.

30. "Discrimination Between Stationary Guassian Processes, Large Sample Results," Will Gersch, January 1977.

31. "Principal Components in the Nonnormal Case: The Test for Sphericity," Christine M. Waternaux, October 1977.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 31 ✓	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): PRINCIPAL COMPONENTS IN THE NONNORMAL CASE: THE TEST FOR SPHERICITY		5. TYPE OF REPORT & PERIOD COVERED: Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): CHRISTINE M. WATERNAUX		8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0442 ✓
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics Stanford University ✓ Stanford, California		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (NR-042-034)
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research Statistics & Probability Program Code 436 Arlington, Virginia 22217		12. REPORT DATE: OCTOBER 1977 13. NUMBER OF PAGES: 21
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report): UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Issued also as Technical Report No. 122 under National Science Foundation Grant MPS 75-09450, Department of Statistics, Stanford University. Also, supported by National Science Foundation Grant MCS 76-09048.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Principal Components. Robustness. Test for Sphericity.

Test for Equality of Variances.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

SEE REVERSE SIDE

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

PRINCIPAL COMPONENTS IN THE NONNORMAL CASE: THE TEST FOR SPHERICITY

Christine M. Waternaux
Harvard University

The limiting distribution of the likelihood ratio statistic W_q for testing the hypothesis of equality of q characteristic roots of a covariance matrix for normal populations is studied for nonnormal populations. It is shown, both theoretically and empirically, that the limiting distribution of W_q is not robust to departures from normality characterized by nonzero fourth cumulants and that W_q cannot be used for these nonnormal populations. For the class of spherically symmetric populations, it is shown that the limiting distribution of W_q is proportional to a chi-square under the null hypothesis of equality of q population roots and to a noncentral chi-square under an appropriate sequence of alternative hypotheses. A corrected test statistic, W_q^*, whose limiting distribution is chi-square, is proposed to test the hypothesis of sphericity for the general class of spherically symmetric populations. Results of a Monte Carlo experiment conducted to compare the performances of W_q and W_q^* are presented for various contaminated normal models at various sample sizes. It is demonstrated that there is little difference between W_q and W_q^* for normal populations, but that dramatic improvements are gained by using W_q^* for the nonnormal populations.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)